Determinantal Point Processes for Mini-Batch Diversification

Cheng Zhang
Disney Research
Pittsburgh, PA, USA
cheng.zhang@disneyresearch.com

Hedvig Kjellström
KTH Royal Institute of Technology
Stockholm, Sweden
hedvig@kth.se

Stephan Mandt
Disney Research
Pittsburgh, PA, USA
stephan.mandt@disneyresearch.com

Abstract

We study a mini-batch diversification scheme for stochastic gradient descent (SGD). While classical SGD relies on uniformly sampling data points to form a mini-batch, we propose a non-uniform sampling scheme based on the Determinantal Point Process (DPP). The DPP relies on a similarity measure between data points and gives low probabilities to mini-batches which contain redundant data, and higher probabilities to mini-batches with more diverse data. This simultaneously balances the data and leads to stochastic gradients with lower variance. We term this approach Diversified Mini-Batch SGD (DM-SGD). We show that regular SGD and a biased version of stratified sampling emerge as special cases. Furthermore, DM-SGD generalizes stratified sampling to cases where no discrete features exist to bin the data into groups. We show experimentally that our method results more interpretable and diverse features in unsupervised setups, and in better classification accuracies in supervised setups.

1 INTRODUCTION

Stochastic gradient descent (SGD) is one of the most important algorithms for scalable machine learning [7, 36, 27]. SGD optimizes an objective function by successively following noisy estimates of its gradient based on mini-batches from a large underlying dataset. We usually assure that this gradient is unbiased, meaning that the expected stochastic gradient equals the true gradient. When combined with a suitably decreasing learning rate schedule, the algorithm converges to a local optimum of the objective [7].

Often we are not interested in learning an unbiased estimator of the gradient, but are rather willing to introduce some bias. There are many reasons for why this might be the case. First, biased SGD schemes such as momentum [34], iterate averaging [40], or preconditioning [9, 16, 43, 46] may reduce the stochastic gradient noise or ease the optimization problem, and therefore often lead to faster convergence. Another reason is that we may decide to actively select samples based on their relevance or difficulty levels such as boosting [10], or because we believe that our dataset is in some respect imbalanced [12]. In this paper, we propose and investigate a biased mini-batch subsampling scheme for imbalanced data.

Real-world data sets are naturally imbalanced. For instance, the sports topic appears more often in the news than biology; the internet contains more images of young people than of senior people, and Youtube has more videos of cats than of bees or ants. Aiming to maximize the probability of generating such training data, machine learning models will refine the dominant information with redundancy but ignore the important but scarce data. For example, a model trained on Youtube data might be very sensitive to different cats but unable to recognize ants. We may therefore decide to try to learn on a more balanced data set by actively selecting diversified mini-batches.

The currently most common tool for mini-batch diversification is stratified sampling [30, 48]. In this approach, one groups the data into a finite set of *strata* based on discrete or continuous features such as a label or cluster assignment. To re-balance the data set, the data can then be subsampled such that each stratum occurs with equal probability in the mini-batch (in the following, we refer to this method as *biased stratified sampling*). Unfortunately, the data are not always amenable to biased stratified sampling because discrete features may not exist, or the data may not be unambiguously clustered. Instead of subsampling based on discrete strata, it would be desirable to diversify the mini-batch based on a soft similarity measure between data points. As we show in this paper, this can be achieved using Determinantal Point Processes (DPPs) [19].

The DPP is a point process which mimics repulsive interactions between samples. Being based on a similarity matrix between the data points, a draw from a DPP yields diversified subsets of the data. The main contribution of this paper is using this mechanism to diversify the mini-batches in stochastic gradient-based learning and analyzing this setup theoretically. In more detail, our main achievements are:

- We present a mini-batch diversification scheme based on DPPs for stochastic gradient algorithms. This approach requires a similarity measure among data points, which can be constructed using low-level features of the data. Since the sampling strategy is independent of the learning objective, diversified mini-batches can be precomputed in parallel and reused for different learning tasks. Our approach applies to both supervised and unsupervised models.
- We prove that our method is a generalization of stratified sampling and i.i.d. mini-batch sampling. Both cases emerge for specific similarity kernels of the data.
- We theoretically analyze the conditions under which the variance of the DM-SGD gradient gets reduced. We also give an unbiased version of DM-SGD which optimizes the original objective without re-balancing the data.
- We carry out extensive experiments on several models and datasets. Our approach leads to faster learning and higher classification accuracies in deep supervised learning. For topic models we find that that the resulting document features are more interpretable and are better suited for subsequent supervised learning tasks.

Our paper is structured as follows. In Section 2 we list related work. Section 3 discusses our main concepts of a diversifed risk, and discuses the DM-SGD method. Section 4 discusses theoretical properties of our approach such as variance reduction. Finally, in Section 5, we give empirical evidence that our approach leads to higher classification accuracy and better feature extractions than i.i.d. sampling.

2 RELATED WORK

We revisit the most relevant prior work based on the following aspects. *Diversification and Stratification* comprises methods which aim at re-balancing the empirical distribution of the data. *Variance reduction* summarizes stochastic gradient methods that aim at faster convergence by reducing the stochastic gradient noise. Finally, we list related applications and extensions of *determinantal point processes*.

Diversification and stratification. Since our method suggests to diversify the mini-batches by of non-uniform subsampling from the data, it relates to stratification methods.

Stratification [30, 29] assumes that the data decomposes into disjoint sub-datasets, called strata. These are formed based on certain criteria such as a class-label. Instead of uniformly sampling from the whole dataset, each stratum is sub-sampled independently, which reduces the variance of the estimator of interest.

Stratified sampling has been suggested as a variance reduction method for stochastic gradient algorithms [11, 48]. If one subsamples the same number of data points from every stratum to form a mini-batch as in [48], one naturally balances the training procedure. This approach was also used in [1]. Our work relates closely to this type of biased stratified sampling. It is different in that it does not rely on discrete strata, but only requires a measure a measure of similarity between data points to achieve a similar effect. This applies more broadly.

Variance reduction. Besides re-balancing the dataset, our approach also reduces the variance of the stochastic gradients. Several ways of variance reduction of stochastic gradient algorithms have been proposed, an important class relying on control variates [26, 32, 35, 44, 38]. A second class of methods relies on non-uniform sampling of mini-batches [8, 11, 33, 39, 48, 49]. None of these methods rely on similarity measures between data points.

Our approach is most closely related to clustering-based sampling (CBS) [11] and stratified sampling (StS) [48]. StS applies stratified sampling to SGD and builds on pre-specified strata. For every stratum, the same number of data points are uniformly selected, and then re-weighted according to the size of the stratum to make the sampling scheme un-biased. CBS uses a similar strategy, but does not require a pre-speficied set of strata. Instead, the strata are formed by pre-clustering the raw data with k-means. (Thus, if the data are clustered based on a class label, CBS is identical to StS.) The problem is that the data are not always amenable to clustering. Second, both StS and CBS ignore the within-cluster variations between data points. In contrast, our approach relies on a continuous measure of similarity between samples. We furthermore show that it is a strict generalization of both setups for particular choices of similarity kernels.

Determinantal point processes. The DPP [19, 25] has been proposed [20, 22, 45] and advanced [4, 23, 24] in the machine learning community in the recent years. It has been applied in subset sampling [18, 24] and results filtering [22].

The DPP has also been used as a diversity-enhancing prior in Bayesian models [20, 45]. In big data setups, the data may overwhelm the prior such that the strength of the prior has to scale with the number of data points; introducing a bias. The approach is furthermore constrained to hierarchical Bayesian models, while our approach applies to all empirical risk minimization problems.

Recently, efficient algorithms have been proposed to make sampling using the DPP more scalable. In the traditional formulation, mini-batch sampling costs $\mathcal{O}(Nk^3)$, with an initial fixed cost of diagonalizing the similarity matrix [19], where N is the size of the data and k is the size of the mini-batch. Recent scalable versions of the DPP rely on core-sets and low-rank approximations and scale more favorably [4, 24]. These versions were used in our large-scale experiments.

3 METHOD

Our method, DM-SGD, uses a version of the DPP for mini-batch sampling in stochastic gradient descent. We show that this balances the underlying data distribution and simultaneously accelerates the convergence due to variance reduction. We briefly revisit DPP first, and then introduce our mini-batch diversification method. Theoretical aspects are then discussed in Section 4.

3.1 DETERMINANTAL POINT PROCESSES

A point process is a collection of points randomly located in some mathematical space. The most prominent example is the Poisson process on the real line [17], which models independently occurring events. In contrast, the DPP [19, 25] models repulsive correlations between these points.

In this paper, we restrict ourselves to a finite set of N points. Denote by $L \in \mathbb{R}^{N\times N}$ a similarity kernel matrix between these points, e.g. based on spatial distances or some other criterion. L is real, symmetric and positive definite, and its elements L_{ij} are some appropriately defined measure of similarity between the i_{th} and j_{th} data. The DPP assigns a probability to subsampling any subset Y of $\{1,\ldots,N\}$, which is proportional to the determinant of the sub-matrix L_Y of L which indexes the subset,

$$\mathscr{P}(Y) = \frac{det(L_Y)}{det(L+I)} \propto det(L_Y). \tag{1}$$

For instance, if $Y = \{i, j\}$ consists of only two elements, then $\mathscr{P}(Y) \propto L_{ii}L_{jj} - L_{ij}L_{ji}$. Because L_{ij} and L_{ji} measure the similarity between elements i and j, being more similar lowers the probability of co-occurrence. On the other hand, when the subset is very diverse, the determinant is bigger and correspondingly its co-occurrence is more likely. The DPP thus naturally diversifies the selection of subsets.

In this paper, we propose to use the DPP to diversify mini-batches. In practice, the mini-batch size is usually constrained by empirical bounds or hardware restrictions. In this case, we want to use DPP conditioned on a given size k. Therefore, a slightly modified version of the DPP is needed, which is called k-DPP [18]. It assigns probabilities to subsets of size k,

$$\mathscr{P}_L^k(Y) = \frac{det(L_Y)}{\sum_{|Y'|=k} det(L_{Y'})}. \tag{2}$$

Apart from conditioning on the size of the subset of points, the k-DPP has the same diversification effect as the DPP [18]. In order to have a fixed mini-batch size we use the k-DPP in this work.

3.2 MINI-BATCH DIVERSIFICATION

The diversifying property of the k-DPP makes it well-suited to diversify mini-batches. We first discuss our learning objective—the diversified risk. We then introduce our algorithm and qualitatively discuss its properties.

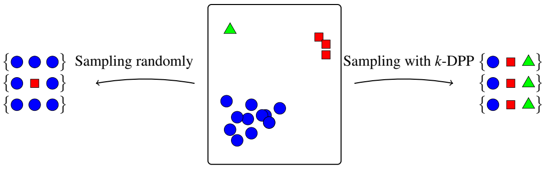


Figure 1: Sampling mini-batches using the k-DPP. For an imbalanced dataset, our method results in diversified mini-batches.

Expected, empirical, and diversified risk. Many problems in machine learning amount to minimizing some loss function $\ell(x,\theta)$ which both depends on a set of parameters θ and on data x. In probabilistic modeling, ℓ could be the negative logarithm of the likelihood of a probabilistic model, or a variational lower bound [6, 14]. We often thereby assume that the data were generated as draws from some underlying unknown data-generating distribution $p_{\text{data}}(x)$, also called the population distribution. To best generalize to unseen data, we would ideally like to minimize this function's expectation under p_{data},

$$J(\theta) = \mathop{\mathbb{E}}_{x\sim p_{\text{data}}}[\ell(x;\theta)] \tag{3}$$

This objective function is also called expected risk [7]. Since $p_{\text{data}}(x)$ is unknown and we believe that our observed data are in some sense a representative draw from the population distribution, we can replace the expectation by an expectation over the *empirical* distribution of the data p_{emp}, which leads to the empirical risk [7],

$$\hat{J}(\theta) = \mathop{\mathbb{E}}_{x\sim p_{\text{emp}}}[\ell(x;\theta)] = \frac{1}{N}\sum_{i=1}^{N}\ell(x_i,\theta). \tag{4}$$

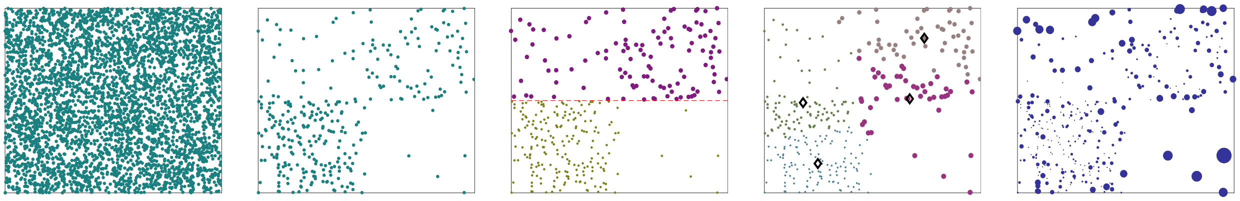
A typical goal in machine learning is not to minimize the empirical risk with high accuracy, but to learn model parameters that generalize well to unseen data. For every data point in a test set, we wish our model to have high predictive accuracy. If this test set is more balanced than the training set (for instance, because it contains all classes to equal proportions in a classification setup), we would naturally like to train our model on a more balanced training set than the original one without throwing away data. In this work, we present a systematic way to achieve this goal based on biased subsampling of the training data. We term the collection of all samples generated from biased subsampling the *balanced dataset*.

To this end, we introduce the *diversified risk*, where we average the loss function over diversified mini-batches $\vec{x}$ of size k,

$$J^*(\theta) = \tfrac{1}{k}\mathop{\mathbb{E}}_{\vec{x}\sim \text{k}-\text{DPP}}[\ell(\vec{x};\theta)], \tag{5}$$

Due to the repulsive nature of k-DPP, similar data points are less likely to co-occur in the same draw. Thus, data points which are very different from the rest are more likely to be sampled and obtain a higher weight, as illustrated in Figure 2 (e).

The diversified risk depends both on the mini-batch size and on the similarity kernel L of the data. A more theoretical analysis of the diversified risk is carried out in Section 4.



(a) True population (b) Dataset (c) Stratified Sampling (d) Pre-clustering (K-means) (e) k-DPP sampling

Figure 2: Visualization of different non-uniform data subsampling schemes on toy data. Panel (a) shows a homogeneous distribution of data. We assume that we only observe an imbalanced subset, shown in panel (b). Panels (c), (d), (e) demonstrate different biased sampling methods that aim at restoring balance in the data. Thicknesses of data points thereby indicate their sampling frequency. Biased stratified sampling (c) relies on dividing the feature space vertically along certain dimensions, whereas pre-clustering (d) defines the strata as clusters obtained from k-means [11] (we used $k = 4$). The black diamonds show the cluster centers and data are colored with respect to their cluster membership. Panel (e) shows the results using the k-DPP, using an RPF kernel of spatial distances as similarity measure between data points. In this example, the k-DPP best restores the balance of the original data set.

Algorithm. Our proposed algorithm directly optimizes the diversified risk in Eq. 5. To this end, we propose SGD updates on diversified mini-batches of fixed size k,

$$\theta_{t+1} = \theta_t - \rho_t \tfrac{1}{k} \sum_{i \in B} \nabla \ell(\theta, x_i), \quad B \sim \mathrm{k-DPP}. \qquad (6)$$

Above, $B \subset \{1, \ldots, N\}$ is a collection of k indices, drawn from the k-DPP. In every stochastic gradient step, we thus sample mini-batches from the k-DPP and carry out an update with decreasing learning rate ρ_t.

Sampling from the k-DPP first requires an eigendecomposition of its kernel. This decomposition can also be approximated and has to be computed only once for one dataset. Drawing a sample then has the computational complexity $\mathcal{O}(Nk^3)$, where k is the mini-batch size, which is much more efficient since k is commonly small. This approach is briefly summarized in Algorithm 1; details on the sampling procedure are given in the supplementary material. For more details, we refer to [19] and to [4, 24] for more efficient sampling procedures.

Algorithm 1 DM-SGD

```
Input: Data $X$, mini-batch size $k$, eigendecomposition $\{(v_n, \lambda_n)\}_{n=1}^N$ of similarity matrix $K$.
for $t = 0$ to MaxIter do
    Sample a mini-batch using the $k$-DPP
    Sample $k$ eigenvectors $V$ using eigenvalues;
    Sample mini-batch $\vec{x}$ of size $k$ using $V$. (See supplement.)
    Update parameters
    $\theta_{t+1} = \theta_t + \rho_t g^*(\theta_t; \vec{x})$ (g* is the gradient estimate)
end
```

Stochastic Variational Inference (SVI) employs SGD for training probabilistic graphical models, such as Latent Dirichlet Allocation (LDA). Every SVI update involves an inner loop. Algorithm 2 shows an application of DM-SGD to SVI for LDA [6]. We thus term it *DM-SVI*.

Algorithm 2 DM-SVI

```
We adopt the notation from [13].
for $t = 0$ to MaxIter do
    Sample a mini-batch using the $k$-DPP;
    Update variational parameters;
    for $j = 0$ to Mini-batch Size do
        Update local variational parameters ( e.g. $\phi$ and $\lambda$ for LDA) for mini-batch.
    end
    Compute the intermediate global parameters as if the mini-batch is replicated $\frac{D}{S}$ times.
    ( e.g. $\tilde{\lambda}_{kw} = \eta + \frac{D}{S} \sum_{s=1}^{S} n_{tw} \phi_{twk}$ for LDA)
    Update the current estimate of the global variational parameters with $\rho_t = (\tau_0 + t)^{-k}$.
    $\lambda = (1 - \rho_t)\lambda + \rho_t \tilde{\lambda}$
end
```

Variance reduction and connections to biased stratified sampling. Dividing the data into different strata and sampling data from each stratum with adjusted probabilities may reduce the variance of SGD. This insight forms the basis of stratified sampling [48], and the related pre-clustering based method [11]. As we will demonstrate rigorously in the next section, our approach also enjoys variance reduction but does not require an artificial partition of the data into clusters.

For many models, the gradient varies smoothly as a function of the data. Subsampling data from diversified regions in data space will therefore decorrelate the gradient contributions. This, in turn, may reduce the variance of the stochastic gradient. To some degree, methods such as biased stratified sampling or pre-clustering sample data from diversified regions, but ignore the fact that gradients *within* clusters may still be highly correlated. If the data are not amenable to clustering, this variance may be just as large as the inter-cluster variance. Our approach does not rely on the notion of clusters. Instead, we have a continuous measure of similarity between samples, given by the similarity kernel. This applies more broadly.

In Figure 2, we investigate how well our subsampling procedure using the k-DPP allows us to recover an original distribution of data from which we only observe an imbalanced subset. Panel (a) shows the original (uniform) distribution of data points, and (b) shows the observed data set which we use to re-estimate the original dataset. While biased stratified sampling (c) or pre-clustering based on k-means (d) need an artificial way of dividing the data into finitely many strata and re-balance their corresponding weights, our approach (e) relies on a continuous similarity measure between data and takes into account both intra-strata and inter-strata variations.

Computational overhead. Sampling from the k-DPP implies a computational overhead over classical SGD. Regarding the overall runtime, the benefits of the approach therefore come mainly into play in setups where each gradient update is expensive. One example is stochastic variational inference for models with local latent variables. For example, in LDA, the computational bottleneck is to update the per-document topic proportions. The time spent on sampling a mini-batch using the k-DPP is only about 10% of the time to infer these local variables and estimate the gradient (See Table 1 in Section 5). Spending this tiny overhead on actively selecting training examples is well invested as the resulting stochastic gradient has a lower variance.

Since the sampling procedure is independent of the learning algorithm, we can parallelize it or even draw the samples as a pre-processing step and reuse them for different hyperparameter settings. Moreover, there are approximate versions of k-DPP sampling which are scalable to big datasets [4, 23]. In this paper, we use the *fast k-DPP* [23] in our large-scale experiments (Section 5.3).

4 THEORETICAL CONSIDERATIONS

In this section, we give the theoretical foundation of the DM-SGD scheme. We first prove that biased stratified sampling and pre-clustering emerge as special cases of our algorithm for particular choices of the kernel matrix L. We then prove that the diversified risk of DM-SGD is a re-weighted variant of the empirical risk, where the weights are given by the marginal likelihoods of the k-DPP (we also present an unbiased DM-SGD scheme which approximates the true gradients, but which performs less favorably in practice). Last, we investigate under which circumstances DM-SGD reduces the variance of the stochastic gradient.

Notation. For what follows, let $m_i \in \{0,1\}$ denote a variable which indicates whether the i_{th} data point was sampled under the k-DPP. Furthermore, let $\mathbb{E}[\cdot] = \mathbb{E}_{m \sim \text{k-DPP}}[\cdot]$ always denote the expectation under the k-DPP. This lets us express the expectation $F(x) = \sum_i f(x_i)$ which depends additively on the data points x_i as

$$\mathbb{E}[\textstyle\sum_{i=1}^N m_i f(x_i)] \equiv \underset{x\sim \text{k-DPP}}{\mathbb{E}}[F(x)] \tag{7}$$

Next, we introduce short hand notations for first and second moments. Denote the marginal probability for a point x_i being sampled as

$$b_i \equiv \mathbb{E}[m_i], \tag{8}$$

which has an analytic form and can be computed efficiently. We also introduce the correlation matrix

$$C_{ij} = \frac{\mathbb{E}[(m_i - b_i)(m_j - b_j)]}{\mathbb{E}[m_i]\,\mathbb{E}[m_j]} = \frac{\mathbb{E}[m_i m_j]}{b_i b_j} - 1. \tag{9}$$

In contrast to minibatch SGD where $\mathbb{E}[m_i m_j] = \mathbb{E}[m_i]\,\mathbb{E}[m_j]$ and hence $C_{ij} = 0$, this is no longer true under the k-DPP. Instead, the correlation can be both negative (when data points are similar) and even positive (when data points are very dissimilar).

Lastly, let $g(\theta, x) = \sum_{i=1}^N g(\theta, x_i)$ denote the gradient of the empirical risk, which is the batch gradient, and $g(\theta, x_i)$ its individual contributions from the data x_i.

We first prove that our algorithm captures two important limiting cases, namely (biased) stratified sampling and pre-clustering.

Proposition 1. Biased stratified sampling (StS) [48], where data from different strata are subsampled with equal probability, is equivalent to DM-SGD with a similarity matrix L, defined as a block-diagonal matrix with

$$L_{ij} = \begin{cases} 1 & H_i = H_j \\ 0 & H_i \neq H_j, \end{cases} \tag{10}$$

where H_i denotes the label for the stratum of data point i.

Proof. It is enough to show that a draw A from the k-DPP which has multiple data points with the same strata assignment has probability zero.

Let $A = a \cup \bar{a}$, where a is a collection of indices which come from the same stratum, and $\bar{a}$ is its disjoint complement. Because of the block-structure of L, we have that

$$\det(L_A) = \det(L_a)\det(L_{\bar{a}}).$$

However, $\det(L_a) = 0$ because it is a matrix of all-ones. Therefore, $\det(L_A) = 0$, and hence A has zero probability under the k-DPP. Therefore, every draw from the k-DPP with L_{ij} defined as above contains at most one data point from each stratum. When k is the same as the number of classes, we recover StS. If k is smaller than the number of classes, we provide a direct generalization of StS. □

Proposition 2. Pre-clustering [11] results as a special case of DM-SGD, with $L_{ij} = 1$ if the data points i and j are assigned to the same cluster, and otherwise $L_{ij} = 0$.

It is furthermore simple to see that regular minibatch SGD results from DM-SGD when choosing the identity kernel.

Next, we analyze the objective function of DM-SGD. We prove that the diversified risk (Eq. 5) is given by a re-weighted version of the empirical risk (Eq. 4) of the data.

Proposition 3. The diversified risk (Eq. 5) can be expressed as a re-weighted empirical risk with the marginal k-DPP weights b_i,

$$J^*(\theta) = \frac{1}{k}\sum_{i=1}^{N} b_i\,\ell(x_i,\theta).$$

As $b_i \to k/N$ in case of a trivial similarity kernel $L = I$, this quantity just becomes the empirical risk.

Proof. We employ the indicators m_i defined above:

$$\begin{aligned} kJ^*(\theta) &= \mathop{\mathbb{E}}_{x\sim \text{kDPP}}[\ell(x;\theta)] = \mathbb{E}[\sum_{i=1}^{N} m_i \ell(x_i;\theta)] \\ &= \sum_{i=1}^{N}\mathbb{E}[m_i]\ell(x_i;\theta) = \sum_{i=1}^{N} b_i \ell(x_i;\theta). \end{aligned}$$

□

The following corollary allows us to construct an unbiased stochastic gradient based on DM-SGD in case we are not interested in re-balancing the population.

Proposition 4. The following SGD scheme leads to an unbiased stochastic gradient:

$$\theta_{t+1} = \theta_t - \rho_t \frac{1}{k}\sum_{i\in B}\frac{1}{b_i}\nabla\ell(\theta,x_i), \quad B \sim \text{kDPP}. \qquad (11)$$

This is a simple consequence of the identity $\mathbb{E}[\sum_{i=1}^{N}\frac{m_i}{b_i}g(\theta;x_i)] = \sum_{i=1}^{N}\mathbb{E}[\frac{m_i}{b_i}]g(\theta;x_i) = g(\theta,x)$.

Finally, we investigate under which circumstances the DM-SGD gradient has a lower variance compared to simple mini-batch SGD on the diversified risk. To this end, consider the gradient components $g(x_i,\theta)$, $g(x_i,\theta)$ of data points i and j, respectively, as well as their correlation C_{ij} under the k-DPP. A sufficient condition for BN-SGD to reduce the variance is given as follows.

Theorem 1. Assume that for all data points x_i and x_j and for all parameters θ in a region of interest, the scalar product $g(x_i,\theta)^\top g(x_j,\theta)$ is always positive (negative) whenever the correlation C_{ij} is negative (positive), respectively, i.e.

$$\forall_{i\neq j} : C_{ij}\, g(x_i,\theta)^\top g(x_j,\theta) < 0. \qquad (12)$$

Then, DM-SGD has a lower variance than SGD.

Remark. The sufficient conditions outlined in Theorem 1 are very strong, but its proof provides us with valuable insights of why variance reduction occurs.

Proof. To begin with, define

$$g^F(\theta,x) = \tfrac{1}{k}\textstyle\sum_{i=1}^{N} b_i\, g(\theta,x_i), \qquad (13)$$

$$g^*(\theta,x) = \tfrac{1}{k}\textstyle\sum_{i=1}^{N} m_i\, g(\theta,x_i), \qquad (14)$$

where g^* is the DM-SGD gradient and $g^F = \mathbb{E}[g^*]$ is the full gradient of the diversified risk.

We denote the difference between the expected and stochastic gradient as

$$\Delta g = g^* - g^F = \tfrac{1}{k}\textstyle\sum_{i=1}^{N}(b_i - m_i)\, g(\theta,x_i), \qquad (15)$$

By construction, this quantity has expectation zero. We are interested in the trace of the stochastic gradient covariance,

$$Var(g^*) = \mathrm{Tr}\,(Cov(g^*)) = \mathbb{E}[\Delta g^\top \Delta g]. \qquad (16)$$

This quantity can be expressed as

$$Var(g^*) = \frac{1}{k^2}\sum_{i,j=1}^{N}\underbrace{\mathbb{E}[(m_i - b_i)(m_j - b_j)]}_{\mathbb{E}[m_i m_j] - b_i b_j} g(x_i,\theta)^\top g(x_j,\theta)$$

We can furthermore compute

$$\begin{aligned} \mathbb{E}[m_i m_j] &= \mathbb{E}[m_i^2]\delta_{ij} + \mathbb{E}[m_i m_j](1-\delta_{ij}) \\ &= \mathbb{E}[m_i]\delta_{ij} + (C_{ij}+1)b_i b_j(1-\delta_{ij}), \end{aligned}$$

where δ_{ij} is the Kronecker symbol (we used $m_i^2 = m_i$).

Collecting all terms, the variance can be written as

$$\begin{aligned} Var(g^*) \;=\; & \frac{1}{k^2}\sum_{i=1}^{N}(b_i - b_i^2)\,||g(x_i,\theta)||_2^2 \\ & + \frac{1}{k^2}\sum_{i\neq j} C_{ij} b_i b_j g(x_i,\theta)^\top g(x_j,\theta). \end{aligned}$$

The first term is just the variance of regular mini-batch SGD, where we sample each data point with probability proportional to b_i, which also optimizes the diversified risk. This term is always positive because $b_i < 1$ and thus $b_i > b_i^2$.

The second term can be both positive and negative. By a clever choice of similarity kernel and resulting correlation function C_{ij} (as defined in Eq. 9), the second term may therefore reduce the variance. We immediately see that this condition exactly corresponds to Eq. 12. This proves our claim. □

Discussion of Theorem 1. If the similarity kernel L relies on spatial distances, nearby data points x_i and x_j have a negative correlation C_{ij} under the k-DPP. However, if the loss function is smooth, $g(x_i,\theta)$ and $g(x_j,\theta)$ tend to align (i.e. have a positive scalar product). Eq. 12 is therefore naturally satisfied for these points. C_{ij} can also be positive: since some combinations of data points are less likely to co-occur, others must be more likely to co-occur. Since these points tend to be far apart, it is reasonable to assume that their gradients show no tendency to align. It is therefore plausible to assume that for these points, Eq. 12 also applies[1].

To summarize, if the condition in Eq. 12 is met, we can guarantee variance reduction relative to mini-batch SGD, and we have given arguments why it is plausible that these are met to some degree when using DM-SGD with a distance-dependent similarity kernel. In our experimental section we show that DM-SGD has a faster learning curve, which we attribute to this phenomenon.

[1] We only need to assure that the negative contributions outweigh the positive ones to see variance reduction.

5 EXPERIMENTS

We evaluate the performance of our method in different settings. In Section 5.1 we demonstrate the usage of DM-SGD for Latent Dirichlet Allocation (LDA) [6], an unsupervised probabilistic topic model. We show that the learned diversified topic representations are better suited for subsequent text classification. In Section 5.2 we evaluate the supervised scenario based on multinomial (softmax) logistic regression with imbalanced data. We compare against stratified sampling, which emerges naturally in this example. In section 5.3 we show that our method also maintains performance on the balanced MNIST data set, where we tested convolutional neural networks. In all the experiments, we pre-sample the mini-batch indices using the k-DPP implementation from [19] for small datasets, and from [23] for big datasets. In this way, sampling is treated as a pre-scheduling step and can easily be parallelized. We found that our approach finds more diversified feature representations (in unsupervised setups) and higher predictive accuracies (in supervised setups). We also found that the k-DPP converges within fewer passes through the data compared to standard minibatch sampling due to variance reduction.

5.1 TOPIC LEARNING WITH LDA

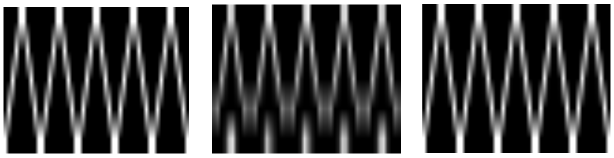
We follow Algorithm 2 for LDA. Firstly, we demonstrate the performance of DM-SVI on synthetic data with LDA. We show that by balancing our mini-batches, we find a much better recovery of the topics that were used to generate the data. Second, we use a real-world news dataset. We demonstrate that we can learn more diverse topics that are also better features for text classification tasks.

In this setting, stratified sampling is not applicable since there is no discrete feature such as a class label available. With only word frequencies available, no simple criterion can be used to divide the data into meaningful strata.

5.1.1 SYNTHETIC DATA

We generate a synthetic dataset (shown in the supplementary material) following the generative process of LDA with a fixed global latent parameter (the graphical topics). We choose distinct patterns as shown in Figure 3 (a), where each row represents a topic and each column represents a word. To generate an imbalanced data set, we use different Dirichlet priors for the per document topic distribution θ. 300 documents are generated with prior (0.5 0.5 0.01 0.01 0.01); 50 with prior (0.01 0.5 0.5 0.5 0.01) and 10 with prior (0.01 0.01 0.01 0.5 0.5). Hence, the first two topics are used very often in the corpus. Topic 3 and 4 are shown a few times and topic 5 appears very rarely.

We fit LDA to recover the topics of the synthetic data using traditional SVI and our proposed DM-SVI respectively. Here, the raw data occurence x is used to construct the similarity matrix $L = xx^T$. We check how well the global parameters are recovered. Fully recovered latent variables indicate that the model is able to capture its underlying structure of the data. Figure 3 (b) shows the estimated per topic words distribution with SVI and Figure 3 (c) shows the result with our proposed DM-SVI.



(a) Ground Truth (b) Est w. SVI LDA (c) Est w. **DM-SVI**

Figure 3: Per topic word distribution for the synthetic data. Each row presents a topic and each column presents a word. (a) shows the ground truth with which the synthetic data is generated using LDA. (b) shows the estimation of this latent variable with LDA using traditional stochastic variational inference (SVI). (c) shows the estimation of this latent variable with DM-SVI

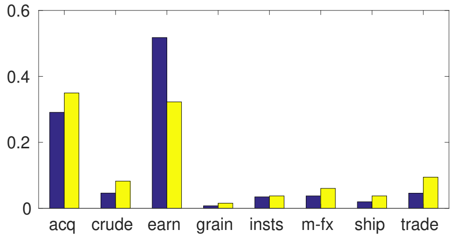


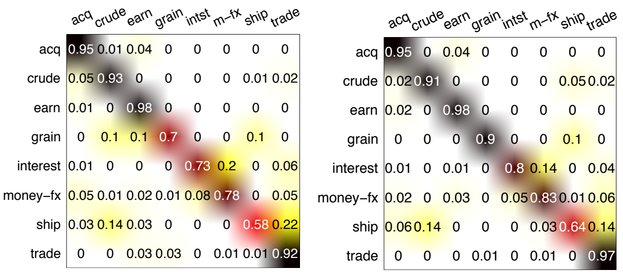
Figure 4: The frequency of class labels of the training dataset (in blue) and of the balanced dataset (in yellow). While explicit class label information is withheld, the algorithm partially balances class contributions.

In Figure 3 (b), we see that the first three topics are recovered using traditional SVI. Topic four is roughly recovered but with information from topic five mixed in. The last topic is not recovered at all, instead, it is a repetition of the first topic. This shows the drawback of the traditional method: when the data is not balanced, the model creates redundant topics to refine the likelihood of the dense data but ignores the scarce data even when they carry important information. In Figure 3 (c), we see that all the topics are correctly recovered thanks to the balanced dataset.

5.1.2 R8 NEWS DATA EXPERIMENT

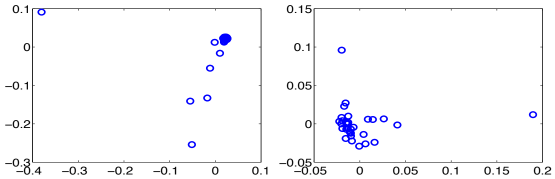
We also evaluate the effect of DM-SVI on the Reuters news R8 dataset [3]. This dataset contains eight classes of news with an extremely imbalanced number of documents per class, as shown in Figure 4 (a). To measure similarities between documents, we represent each document with a vector x of the tf-idf [37] scores of each word in the document. Then define an annealed linear kernel $L(x_i, x_j) = x_i^\rho x_j^\rho$ with parameter $\rho = 0.1$, which is more sensitive to small feature overlap. We run LDA with SVI and DM-SVI with one effective pass through the data, where we set the mini-batch size to 80 and use 30 topics.

We first compare the frequencies at which documents with particular labels were sub-sampled. While Figure 4 shows



(a) CM w.SVI (b) CM w.**DM-SVI**.

Figure 5: Confusion matrix for text classification based on LDA features obtained from SVI (a) and the proposed DM-SVI (b). DM-SVI features lead to better accuracies.



(a) First 2 PC w. SVI (b) First 2 PC w. **DM-SVI**

Figure 6: First 2 principle components of topic word distributions. DM-SVI topic vectors (b) are more diverse compared to SVI (a).

the actual frequency of these classes in the original data set compared with the frequency of these classes over the balanced dataset (a collection of sampled mini-batches using the k-DPP). We can see that the number of documents is more balanced among different classes.

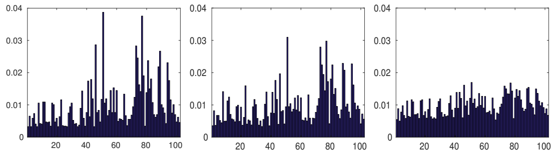
To demonstrate that DM-SVI leads to a more useful topic representation, we classify each document in the test-set based on the learned topic proportions with a linear SVM. The global variable (per-topic word distribution) is only trained on the training set. The resulting confusion matrices are shown in Figure 5 using traditional SVI and DM-SVI respectively. With traditional SVI, the average performance over 8 classes is 82.11%; the total accuracy (number of correctly classified documents over number of test documents) is 94.11%. With DM-SVI, the average performance over 8 classes is 87.24% and the total accuracy is 94.7%.

Thus the overall classification performance is improved using DM-SVI features, and especially the performance on the classes with few documents (such as "grain" and "ship") is improved significantly.

We also visualize the first two principal components (PC) of the the global topics in Figure 6. In traditional SVI, many topics are redundant and share large parts of their vocabulary, resulting in a single dense cluster. In contrast, we see that the topics in DM-SVI are more spread out. In this regard, DM-SVI achieves a similar effect as when using diversity priors as in [20] without the need to grow the prior with the data. The top words from each topic are shown in the appendix, where we present more evidence that the topics learned by DM-SVI are more diverse.

Size	k = 10	k =30	k=50	k=80
Relative cost	0.114%	1.097%	3.191%	8.971%

Table 1: LDA on the R8 dataset. Relative cost of mini-batch sampling as a fraction of the cost of a gradient update. Different values of mini-batch size k are shown.



(a) Labels (b) $w = 0.3$ (c) $w = 0.9$

Figure 7: The frequency of data in each class of the training dataset and of the balanced dataset using different weights w (Eq. 17). There are 102 classes of flowers in total and each bar presents the percentage of data belongs to one class. Minibatch size 50 is used as an example here for (b) and (c).

The relative costs of sampling per iteration for LDA is shown in Table 1. Because every local update is expensive, the relative overhead of mini-batch sampling is small. More details are given in the appendix.

5.2 MULTICLASS LOGISTIC REGRESSION

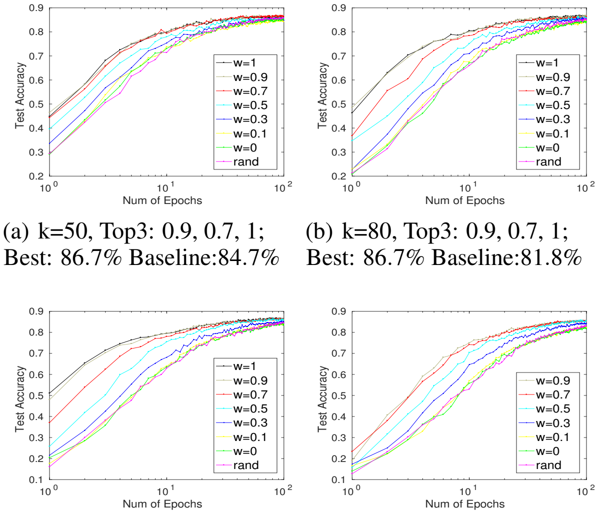
In this section, we demonstrate DM-SGD on a fine-grained classification task. The Oxford 102 flower dataset [31, 41] is used here for evaluation.

Many datasets in computer vision are balanced even though the true collected dataset is extremely imbalanced. The true reason is that the performance of machine learning models usually suffer from imbalanced training data. One example is the Oxford 102 flower dataset which contains 1020 images in the training set with 10 images per class. However, in the test set, 6149 images are available with high imbalance. In this experiment, we make the learning task harder. We use the original testing set for training and use the original training set for testing. This setting demonstrates the real life scenario where we only can collect data with bias but wish the model to perform well in all different situations. Off-the-shelf CNN features [41] are used in this experiment. A pre-trained VGG16 network [42] is used for the feature extraction. We use the first fully connected layer as features, since [5] shows that this layer is most robust.

The similarity kernel L of the k-DPP was constructed as follows. We chose a linear kernel $L = FF^\top$, where F is a weighted concatenation of the fc1 features X_{fc1} and the labels a one-hot-vector representation of the class label H,

$$F = [(1-w)X_{fc1} \;\; wH], \quad 0 \le w \le 1. \tag{17}$$

This kernel construction enables the population to be balanced both among classes and within classes. When w is large, the algorithm focuses more on the class labels. When w is small, balancing is performed mostly based on the fea-



(a) k=50, Top3: 0.9, 0.7, 1; Best: 86.7% Baseline:84.7%

(b) k=80, Top3: 0.9, 0.7, 1; Best: 86.7% Baseline:81.8%

(c) k=102, Top3: 1, 0.9, 0.7; Best: 86.5% Baseline:84.5%

(d) k=150, Top3: 0.7, 0.5, 0.9; Best: 85.5% Baseline:83.1%

Figure 8: Test accuracy as a function of training epochs on the Oxford 102 multi-class classification task. We show DM-SGD for different values of w, with $w = 1$ being biased stratified sampling (see Eq. 17 and the discussion below). The plot caption indicates the batch size k and the three best performing values of w. 'Rand' indicates regular SGD sampling. We listed the final test accuracy after convergence, where "Best" indicates the best performance within our DM-SGD experiments, and "Baseline" indicates regular SGD as our baseline. The improvement is up to 5%.

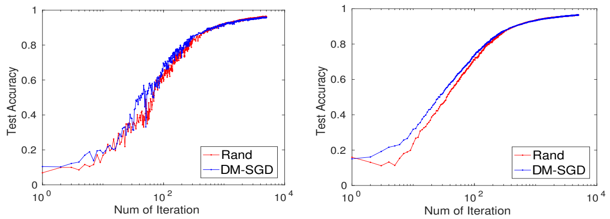
tures. The weighting factor w is a free parameter. As $w = 1$ results in stratified sampling (see Theorem 1), this baseline is naturally captured in our approach.

In this setting, the class label is a natural criterion to divide the data into strata. One can then re-sample the same amount of data from each stratum in order to re-balance the data set. Such a mechanism constrains the mini-batch size to be $k = sM$ where M is the number of classes/strata and s is a positive integer. As proved in Section 4, when $k = M$ and $w = 1$, DM-SGD is equivalent to this type of (biased) stratified sampling.

Figure 7 shows the percentage of data in each class for the original dataset and with the balanced dataset. It shows that with larger w, the dataset is more balanced among classes. More examples are shown in the supplementary material.

We demonstrate this application with a standard linear Softmax classifier for multi-class classification. In our case, the inputs are the off-the-shelf CNN fc1 feature (X_{fc1}). We can also view this procedure as fine-tuning a neural network.

Figure 8 shows how the test accuracy changes with respect to each training epoch. We compare the DM-SGD with different weights against random sampling. The learning rate schedule is kept the same among different experiments. Different mini-batch sizes k are used, which is shown in the caption of each panel in the figure. We can see that with DM-SGD, we can reach a high model performance more rapidly. Additionally, for a classification task, balancing data with respect to classes is important since the performance is better in general for bigger w. On the other hand, the feature information is essential as well since the best performance is mostly obtained with $w = 0.9$ and $w = 0.7$. Comparing these plots, we can see that the performance benefits more when the mini-batch size is comparably small. Small mini-batches in general are preferred due to low cost and our method can maximize the usage of small mini-batches.



(a) k=10

(b) k=200

Figure 9: Same quantities shown as in Fig. 8, but for the MNIST data set, which is more balanced.

5.3 CNN CLASSIFICATION ON MNIST

Finally, we show the performance of our method in a scenario where the dataset is balanced, which is less preferable scenario for DM-SGD. Here we consider the MNIST dataset [21], which contains approximately the same number of examples per hand-written digits. Since our method is independent of the model, we can use any low level data statistics. Here, we demonstrate DM-SGD with raw data features and apply it to training a CNN. Here, we construct the similarity kernel using a RBF kernel. For the low level feature, we use the normalized raw pixel value X directly. To encode both class information and label information, we use $F = [(1-w)X \quad wH]$ to compute the similarities matrix, where $w = 0.5$ for this experiment. We use half of the training data from MNIST to train a 5-layer CNN as in [2]. Figure 9 shows the test accuracy from each iteration with mini-batch size 10 and 200 respectively. We can see that even if the data are balanced, DM-SGD still performs better than random sampling due to its variance reduction property.

6 CONCLUSION

We proposed a diversified mini-batch sampling scheme based on determinantal point processes. Our method, DM-SGD, builds on a similarity matrix between the data points and suppresses the co-occurance of similar data points in the same mini-batch. This leads to a training outcome which generalizes better to unseen data. We also derived sufficient conditions under which the method reduces the variance of the stochastic gradient, leading to faster learning. We showed that our approach generalizes both stratified sampling and pre-clustering. In the future, we will explore the possibility to further improve the efficiency of the algorithm with data reweighing [28] and tackle imbalanced learning problems involving different modalities for supervised [47] and multi-modal [15] settings.

References

[1] Image classification with imagenet. `https://github.com/soumith/imagenet-multiGPU.torch/blob/master/dataset.lua`.

[2] Multilayer convolutional network. `https://www.tensorflow.org/get_started/mnist/pros`.

[3] R8 dataset. `http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html`.

[4] R. H. Affandi, A. Kulesza, E. B. Fox, and B. Taskar. Nystrom approximation for large-scale determinantal processes. In *AISTATS*, pages 85–98, 2013.

[5] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson. From generic to specific deep representations for visual recognition. In *CVPR WS*, pages 36–45, 2015.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, 2003.

[7] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–186. 2010.

[8] D. Csiba and P. Richtarik. Importance sampling for minibatches. *arXiv:1602.02283*, 2016.

[9] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12(Jul):2121–2159, 2011.

[10] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*, pages 23–37. Springer, 1995.

[11] T.F Fu and Z.H. Zhang. CPSG-MCMC: Clustering-based preprocessing method for stochastic gradient MCMC. In *AISTATS*, 2017.

[12] H. He and E. A. Garcia. Learning from imbalanced data. *TKDE*, 21(9):1263–1284, 2009.

[13] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *NIPS*, pages 856–864, 2010.

[14] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[15] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, 2016.

[16] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[17] J. F. C. Kingman. *Poisson processes*. Wiley Online Library, 1993.

[18] A. Kulesza and B. Taskar. k-DPPs: Fixed-size determinantal point processes. In *ICML*, pages 1193–1200, 2011.

[19] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv:1207.6083*, 2012.

[20] J. T. Kwok and R. P. Adams. Priors for diversity in generative latent variable models. In *NIPS*, pages 2996–3004, 2012.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] D. Lee, G. Cha, M. H. Yang, and S. Oh. Individualness and determinantal point processes for pedestrian detection. In *ECCV*, pages 330–346, 2016.

[23] C. T. Li, S. Jegelka, and S. Sra. Fast DPP sampling for nyströom with application to kernel methods. *arXiv:1603.06052*, 2016.

[24] C.T. Li, S. Jegelka, and S. Sra. Efficient sampling for k-determinantal point processes. *arXiv:1509.01618*.

[25] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*.

[26] S. Mandt and D. M. Blei. Smoothed gradients for stochastic variational inference. In *NIPS*, pages 2438–2446, 2014.

[27] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *arXiv:1704.04289*, 2017.

[28] S. Mandt, J. McInerney, F. Abrol, R. Ranganath, and D. M. Blei. Variational Tempering. In *AISTATS*, pages 704–712, 2016.

[29] M. D. McKay, R. J. Beckman, and W. J. Conover. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[30] J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

[31] M. E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[32] J. Paisley, D. Blei, and M. Jordan. Variational bayesian inference with stochastic search. *arXiv:1206.6430*, 2012.

[33] D. Perekrestenko, V. Cevher, and M. Jaggi. Faster coordinate descent via adaptive importance sampling. In *AISTATS*, 2017.

[34] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[35] R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *AISTATS*, pages 814–822, 2014.

[36] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Herbert Robbins Selected Papers*, pages 111–135. Springer, 1985.

[37] S. Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, 60(5):503–520, 2004.

[38] T. Salimans and D.A. Knowles. On using control variates with stochastic approximation for variational bayes and its connection to stochastic linear regression. *arXiv:1401.1022*.

[39] M. Schmidt, R. Babanezhad, M. O. Ahmed, A. Defazio, A. Clifton, and A. Sarkar. Non-uniform stochastic average gradient method for training conditional random fields. In *AISTATS*, 2015.

[40] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, pages 1–30, 2013.

[41] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR WS*, pages 806–813, 2014.

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[43] T. Tieleman and G. Hinton. Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*.

[44] C. Wang, X. Chen, A. J. Smola, and E. P. Xing. Variance reduction for stochastic gradient optimization. In *NIPS*, pages 181–189, 2013.

[45] P.T. Xie, Y.T. Deng, and E. Xing. Diversifying restricted boltzmann machine for document modeling. In *ACM SIGKDD*, pages 1315–1324. ACM, 2015.

[46] M. D. Zeiler. ADADELTA: an adaptive learning rate method. *arXiv:1212.5701*, 2012.

[47] C. Zhang and H. Kjellström. How to Supervise Topic Models. In *ECCV WS*, 2014.

[48] P.L. Zhao and T. Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. *arXiv:1405.3080*.

[49] P.L. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *ICML*, pages 1–9, 2015.